

Mutual Information, Neural Networks and the Renormalization Group

Maciej Koch-Janusz¹ and Zohar Ringel²

¹*Institute for Theoretical Physics, ETH Zurich, 8093 Zurich, Switzerland*

²*Rudolf Peierls Centre for Theoretical Physics, Oxford University, Oxford OX1 3NP, United Kingdom*

Physical systems differing in their microscopic details often display strikingly similar behaviour when probed at low energies. Those universal properties, largely determining their physical characteristics, are revealed by the powerful renormalization group (RG) procedure, which systematically retains “slow” degrees of freedom and integrates out the rest. However, the important degrees of freedom may be difficult to identify. Here we demonstrate a machine learning (ML) algorithm capable of identifying the relevant degrees of freedom without any prior knowledge about the system. We introduce an artificial neural network based on a model-independent, information-theoretic characterization of a real-space RG procedure, performing this task. We apply the algorithm to classical statistical physics problems in two dimensions.

I. INTRODUCTION

Machine learning has been captivating public attention lately due to groundbreaking advances in automated translation, image and speech recognition [1], game-playing [2], and achieving super-human performance in tasks in which humans excelled while more traditional algorithmic approaches struggled [3]. The applications of ML techniques in physics are very recent, initially leveraging the trademark prowess of ML in classification and pattern recognition and applying them to classify phases of matter [4–8] or exploiting the neural networks’ potential as efficient non-linear approximators of arbitrary functions [9, 10] to introduce a new numerical simulation method for quantum systems [11]. However, the exciting possibility of employing machine learning not as a numerical simulator, or a hypothesis tester, but as an integral part of the physical *reasoning* process is still largely unexplored and, given the staggering pace of progress in the field of artificial intelligence, of fundamental importance and promise.

The renormalization group (RG) approach has been one of the conceptually most profound tools of theoretical physics since its inception. It underlies the seminal work on critical phenomena [12], the discovery of asymptotic freedom in quantum chromodynamics [13], and of the Kosterlitz-Thouless phase transition [14, 15]. The RG is not a monolith, but rather a conceptual framework comprising different techniques: real-space RG [16], functional RG [17], density matrix renormalization group (DMRG) [18], among others. While all those schemes differ quite substantially in details, style and applicability there is an underlying physical intuition which encompasses all of them – the essence of RG lies in identifying the “relevant” degrees of freedom and integrating out the “irrelevant” ones iteratively, thereby arriving at a universal, low-energy effective theory. However potent the RG idea, those relevant degrees of freedom need to be identified first [19, 20]. This is often a challenging conceptual step, particularly for strongly interacting systems and may involve a sequence of mathematical mappings to models, whose behaviour is better understood [21, 22].

Here we introduce an artificial neural network algorithm identifying the physically relevant degrees of freedom in a spatial region. The input data are samples of the probability distribution of the system configurations, no further knowledge about the microscopic details of the system is provided. The internal parameters of the network, which ultimately encode the degrees of freedom of interest, are optimized (‘learned’, in neural networks parlance) by a training algorithm based on evaluating real-space mutual information (RSMI) between spatially separated regions. We validate our approach by studying the Ising and dimer models of classical statistical physics in two dimensions; the robustness of RSMI algorithm to physically irrelevant noise is demonstrated.

The correct distillation of the important degrees of freedom, in the spirit of a real-space RG procedure [16], is not only a crucial technical step – it allows to gain insights about the correct way of thinking about the problem at hand.

II. THE REAL SPACE MUTUAL INFORMATION ALGORITHM

Before going into more detail, let us provide a bird’s eye view of our method and results. We begin by phrasing the problem in probabilistic/information-theoretic terms, an approach also investigated in Refs. [23–27]. To this end, we consider a small “visible” spatial area \mathcal{V} , which together with its environment \mathcal{E} forms the system \mathcal{X} , and we define a particular conditional probability distribution $P_{\Lambda}(\mathcal{H}|\mathcal{V})$, which describes how the relevant degrees of freedom \mathcal{H} (dubbed “hiddens”) in \mathcal{V} depend on both \mathcal{V} and \mathcal{E} . We then show that the sought-after conditional probability distribution is found by an algorithm maximizing an information-theoretic quantity, the mutual information (MI), and that this algorithm lends itself to a natural implementation using artificial neural networks (ANNs). Finally, we provide a verification of our claims by considering two paradigmatic models of statistical physics: the Ising model – for which the RG procedure yields the famous Kadanoff block spins – and the dimer model, whose RG is much less trivial.

Consider then a classical system of local degrees of freedom $\mathcal{X} = \{x_1, \dots, x_N\} \equiv \{x_i\}$, defined by a Hamiltonian energy function $H(\{x_i\})$ and associated statistical probabilities $P(\mathcal{X}) \propto e^{-\beta H(\{x_i\})}$, where β is the inverse temperature. Alternatively (and sufficiently for our purposes), the system is given by Monte Carlo samples of the equilibrium distribution $P(\mathcal{X})$. We denote a small spatial region of interest by $\mathcal{V} \equiv \{v_i\}$ and the remainder of the system by $\mathcal{E} \equiv \{e_i\}$, so that $\mathcal{X} = (\mathcal{V}, \mathcal{E})$. We adopt a probabilistic point of view, and treat \mathcal{X}, \mathcal{E} etc. as random variables. Our goal is to extract the relevant degrees of freedom \mathcal{H} from \mathcal{V} .

“Relevance” is understood here in the following way: the degrees of freedom RG captures govern the long distance behaviour of the theory, and therefore the experimentally measurable physical properties; they carry the most information about the system at large, as opposed to local fluctuations. We thus formally define the random variable \mathcal{H} as a composite function of degrees of freedom in \mathcal{V} maximizing the *mutual information* (MI) [28] between \mathcal{H} and the environment \mathcal{E} .

Mutual information, denoted by I_Λ , measures the *total* amount of information about one random variable contained in the other (thus it is more general than correlation coefficients, which measure monotonic relations between variables, only). It is given in our setting by:

$$I_\Lambda(\mathcal{H} : \mathcal{E}) = \sum_{\mathcal{H}, \mathcal{E}} P_\Lambda(\mathcal{E}, \mathcal{H}) \log \left(\frac{P_\Lambda(\mathcal{E}, \mathcal{H})}{P_\Lambda(\mathcal{H})P(\mathcal{E})} \right), \quad (1)$$

The unknown distribution $P_\Lambda(\mathcal{E}, \mathcal{H})$ and its marginalization $P_\Lambda(\mathcal{H})$, depending on a set of parameters Λ (which we keep generic at this point), are functions of $P(\mathcal{V}, \mathcal{E})$ and of $P_\Lambda(\mathcal{H}|\mathcal{V})$, which is the central object of interest. In the supplementary materials we discuss the relation of this approach to RG to the more standard procedures.

Finding $P_\Lambda(\mathcal{H}|\mathcal{V})$ which maximizes I_Λ under certain constraints is a well-posed mathematical question and has a *formal* solution [29]. Since, however, the space of probability distributions grows exponentially with number of local degrees of freedom, it is in practice impossible to use without further assumptions for any but the smallest physical systems. Our approach is to exploit the remarkable dimensionality reduction properties of artificial neural networks (ANNs) [9]. We use restricted Boltzmann machines (RBM), a class of probabilistic ANNs well adapted to approximating arbitrary data probability distributions. An RBM is composed of two layers of nodes, the “visible” layer, corresponding to local degrees of freedom in our setting, and a “hidden” layer. The interactions between the layers are defined by an energy function $E_\Theta \equiv E_{a,b,\theta}(\mathcal{V}, \mathcal{H}) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{ij} v_i \theta_{ij} h_j$, such that the joint probability distribution for a particular configuration of visible and hidden degrees of freedom is given by a Boltzmann weight:

$$P_\Theta(\mathcal{V}, \mathcal{H}) = \frac{1}{Z} e^{-E_{a,b,\theta}(\mathcal{V}, \mathcal{H})}, \quad (2)$$

with Z the normalization. The goal of training of an ANN is to find parameters θ_{ij} (“weights” or “filters”) and a_i, b_i optimizing a chosen objective function.

Three distinct RBMs are used: two are trained as efficient approximators of the probability distributions $P(\mathcal{V}, \mathcal{E})$ and $P(\mathcal{V})$, using the celebrated contrastive divergence (CD) algorithm [30]. Their trained parameters are used by the third network [see Fig. 1(B)], which has a different objective: to find $P_\Lambda(\mathcal{H}|\mathcal{V})$ maximizing I_Λ , we introduce the real space mutual information (RSMI) network, whose architecture is shown in Fig. 1(A). The hidden units of RSMI correspond to coarse-grained variables \mathcal{H} .

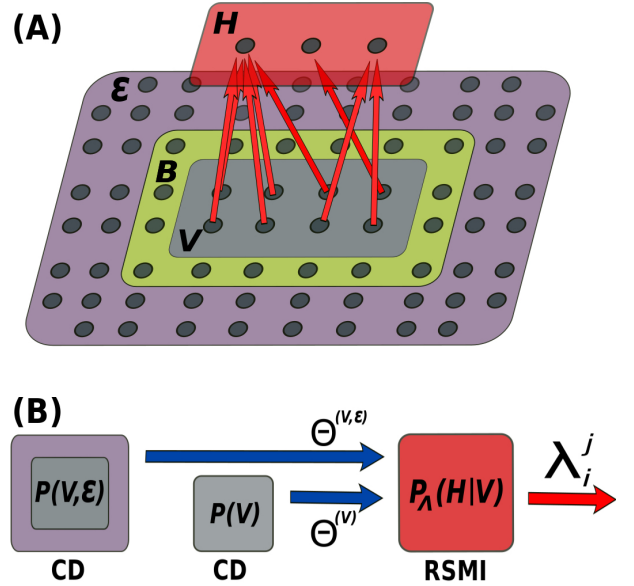


FIG. 1. (A) The RSMI neural network architecture: the hidden layer \mathcal{H} is directly coupled to the visible layer \mathcal{V} via the weights λ_i^j (red arrows), however the training algorithm for the weights estimates MI between \mathcal{H} and the environment \mathcal{E} . The buffer \mathcal{B} , is introduced to filter out local correlations within \mathcal{V} (see supplementary materials). (B) The workflow of the algorithm: the CD-algorithm trained RBMs learn to approximate probability distributions $P(\mathcal{V}, \mathcal{E})$ and $P(\mathcal{V})$. Their final parameters, denoted collectively by $\Theta^{(\mathcal{V}, \mathcal{E})}$ and $\Theta^{(\mathcal{V})}$, are inputs for the main RSMI network learning to extract $P_\Lambda(\mathcal{H}|\mathcal{V})$ by maximizing I_Λ . The final weights λ_i^j of the RSMI network identify the relevant degrees of freedom. For Ising and dimer problems they are shown in Figs. 2 and 4.

The parameters $\Lambda = (a_i, b_j, \lambda_i^j)$ of the RSMI network are trained by an iterative procedure. At each iteration a Monte Carlo estimate of function $I_\Lambda(\mathcal{H} : \mathcal{E})$ and its gradients is performed for the current values of parameters Λ . The gradients are then used to improve the values of weights in the next step, using a stochastic gradient descent procedure.

III. RESULTS

To validate our approach we consider two important classical models of statistical physics: the (critical) Ising model, whose RG is simpler, since the coarse-grained degrees of freedom resemble the original ones, and the fully-packed dimer model, where they are entirely different.

The Ising Hamiltonian on a two-dimensional square lattice is:

$$H_I = \sum_{\langle i,j \rangle} s_i s_j, \quad (3)$$

with $s_i = \pm 1$ and the summation over nearest neighbours. Real-space RG of the Ising model proceeds by the block-spin construction [16], whereby each 2×2 block of spins is coarse grained into a single effective spin, whose orientation is decided by a “majority rule”.

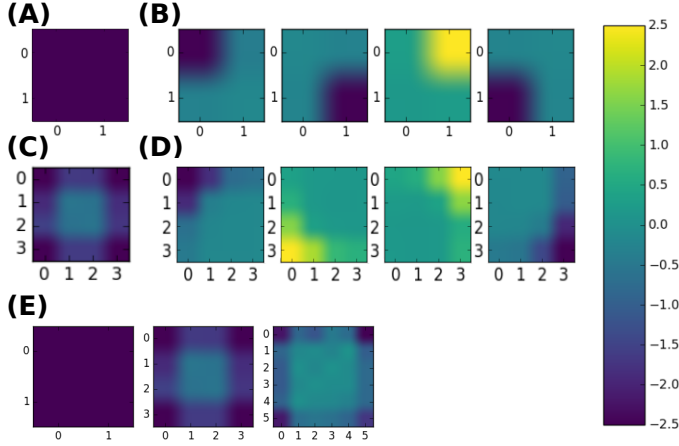


FIG. 2. The weights of the RSMI network trained on Ising model. The ANN couples strongly to areas with large absolute value of the weights. (A) $N_h = 1$ hidden neuron, \mathcal{V} area of 2×2 spins: the ANN discovers Kadanoff blocking (B) $N_h = 4$ for 2×2 area \mathcal{V} . (C) $N_h = 1$ for a 4×4 visible area. (D) $N_h = 4$ for a 4×4 visible area. (E) Comparison of $N_h = 1$ weights for area size of 2×2 , 4×4 , 6×6 – a boundary coupling behaviour (discussed in Supplementary Materials) may be observed.

The results of the RSMI algorithm trained on Ising model samples are shown in Fig. 2. We vary the number of both hidden neurons N_h and the visible units, which are arranged in a 2D area \mathcal{V} of size $L \times L$ [see Fig. 1(A)]. For a 4 spin area the network indeed rediscovers the famous Kadanoff block-spin: Fig. 2(A) shows a single hidden unit coupling uniformly to 4 visible spins, i.e. the orientation of the hidden unit is decided by the average magnetisation in the area. Fig. 2(B) is a trivial but important sanity check: given 4 hidden units to extract relevant degrees of freedom from an area of 4 spins, the network couples each hidden unit to a different spin, as expected.

We also compare the weights for areas \mathcal{V} of different size, which are generalizations of Kadanoff procedure to

larger blocks. We find the network couples to the boundaries of the area \mathcal{V} to maximize MI with the rest of the system [see Figs. 2(C,D,E)]. This physical insight the RSMI provides can in fact be shown to hold exactly for a standard real-space RG that in the limit of number of coarse grained variables equaling the size of the boundary (see supplementary materials).

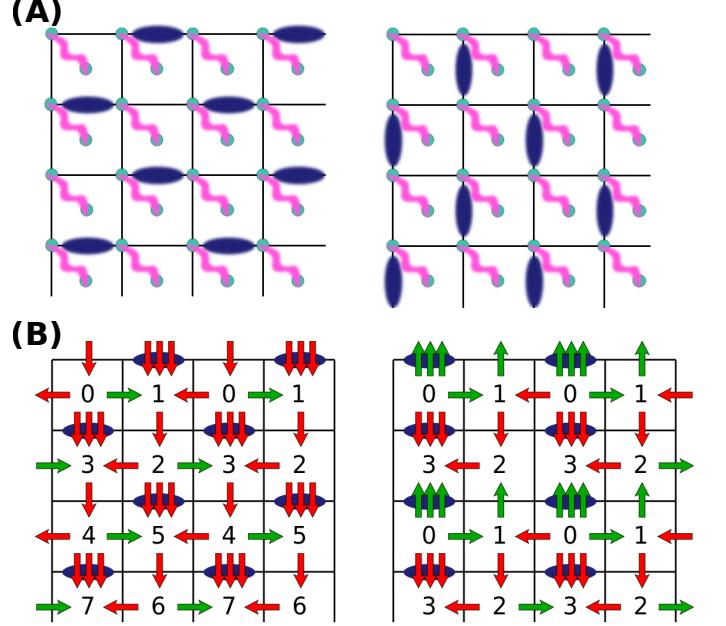


FIG. 3. (A) Two sample dimer configurations (blue links), corresponding to E_y and E_x electrical fields, respectively. The coupled pairs of additional spin degrees of freedom on vertices and faces of the lattice (wiggly lines) are decoupled from the dimers and from each other. Their fluctuations constitute irrelevant noise. (B) An example of mapping the dimer model to local electric fields. The so-called staggered configuration on the left maps to uniform nonvanishing field in the vertical direction: $\langle E_y \rangle \neq 0$. The “columnar” configuration on the right produces both E_x and E_y which are zero on average (see Ref. [31] for details of the mapping).

We next study the dimer model, given by an entropy-only partition function, which counts the number of dimer coverings of the lattice, i.e. subsets of edges such that every vertex is the endpoint of exactly one edge. Fig. 3(A) shows sample dimer configurations (and additional spin degrees of freedom added to generate noise). This deceptively simple description hides nontrivial physics [32] and correspondingly, the RG procedure for the dimer model is more subtle, since – contrary to the Ising case – the correct degrees of freedom to perform RG on are not dimers, but rather look like effective local electric fields. This is revealed by a mathematical mapping to a “height field” h (see Figs.3(A,B) and Ref. [31]), whose gradients behave like electric fields. The continuum limit of the dimer model is given by the following action:

$$S_{dim}[h] = \int d^2x (\nabla h(\vec{x}))^2 \equiv \int d^2x \vec{E}^2(\vec{x}), \quad (4)$$

and therefore the coarse-grained degrees of freedom are low-momentum (Fourier) components of the electrical fields E_x, E_y in the x and y directions. They correspond to “staggered” dimer configurations shown in Fig. 3(A).

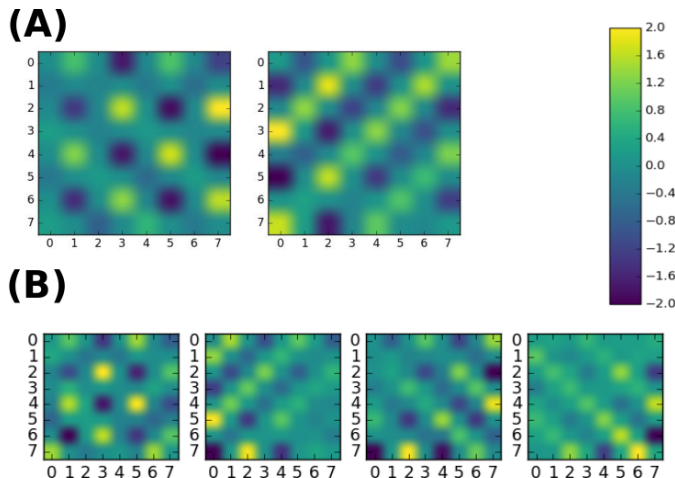


FIG. 4. The weights of the RSMI network trained on dimer model data: (A) $N_h = 2$ hidden neurons for a visible area \mathcal{V} of 8×8 spins. The two filters recognize E_y and $E_x + E_y$ electrical fields, respectively [compare with dimer patterns in Fig. 3(A)]. (B) The trained weights for $N_h = 4$ hidden neurons.

Remarkably, the RSMI algorithm extracts the local electric fields from the dimer model samples without any knowledge of those mappings. In Fig. 4 the weights for $N_h = 2$ and $N_h = 4$ hidden neurons, for an 8×8 area [similar to Fig. 3(A)] are shown: the pattern of large negative (blue) weights couples strongly to a dimer pattern corresponding to local uniform E_y field [see left pannels of Figs. 3(A,B)]. The large positive (yellow) weights select an identical pattern, translated by one link. The remaining neurons extract linear superpositions $E_x + E_y$ or $E_x - E_y$ of the fields.

To demonstrate the robustness of the RSMI, we added physically irrelevant noise, forming nevertheless a pronounced pattern, which we model by additional spin de-

grees of freedom, strongly coupled (ferromagnetically) in pairs [wiggly lines in Fig. 3(A)]. Decoupled from the dimers, and from other pairs, they form a trivial system, whose fluctuations are short-range noise on top of the dimer model. Vanishing weights [green in Figs. 4(A,B)] on sites where pairs of spins reside prove RSMI discards their fluctuations as irrelevant for long-range physics, despite their regular pattern.

IV. OUTLOOK

Artificial neural networks based on real-space mutual information optimization have proven capable of extracting complex information about physically relevant degrees of freedom. This approach is an example of a new paradigm in applying ML in physics: the internal data representations discovered by suitably designed ML systems are not just technical means to an end, but instead are a clear reflection of the underlying structure of the physical system (see also [33]). In spite of its “black box” reputation, the innards of ML architecture may teach us fundamental lessons. This raises the prospect of employing machine learning in science in a collaborative fashion, exploiting the machines’ power to distill subtle information from vast data, and human creativity and background knowledge [34].

Numerous further research directions can be pursued. Most directly, equilibrium systems with less understood relevant degrees of freedom – *e.g.* disordered and glassy systems – can be investigated. Furthermore, though we applied our algorithm to classical systems, the extension to quantum domain is possible via the quantum-to-classical mapping of Euclidean path integral formalism. We envisage extending RSMI to a full RG scheme, *i.e.* using additional ANNs to extract the effective Hamiltonian of the coarse-grained degrees of freedom and possibly reconstruct the RG flow. To this end a formal analysis of the mutual-information based RG procedure may prove fruitful, also from theory perspective. Finally, applications of RSMI beyond physics are possible, since it offers an ANN implementation of a variant of Information Bottleneck method [29], succesful in compression and clustering analyses [35].

-
- [1] Y. LeCun, Y. Bengio, and Hinton G.E., “Deep learning,” *Nature* **521**, 436–444 (2015).
 - [2] D. Silver and et al., “Mastering the game of Go with deep neural networks and tree search,” *Nature* **529**, 584–589 (2016).
 - [3] John R. Hershey, Steven J. Rennie, Peder A. Olsen, and Trausti T. Kristjansson, “Super-human multi-talker speech recognition: A graphical modeling approach,” *Comput. Speech Lang.* **24**, 45–66 (2010).
 - [4] J. Carrasquilla and R. G. Melko, “Machine learning phases of matter,” *Nature Physics* (2017).
 - [5] G. Torlai and R. G. Melko, “Learning thermodynamics with Boltzmann machines,” *Phys. Rev. B* **94**, 165134 (2016).
 - [6] E. P. L. van Nieuwenburg, Y.-H. Liu, and S. D. Huber, “Learning phase transitions by confusion,” *Nature Physics* (2016).
 - [7] L. Wang, “Discovering phase transitions with unsupervised learning,” *Phys. Rev. B* **94**, 195105 (2016).
 - [8] T. Ohtsuki and T. Ohtsuki, “Deep Learning the Quantum Phase Transitions in Random Electron Systems: Applications to Three Dimensions,” *Journal of the Physical*

- Society of Japan **86**, 044708 (2017).
- [9] G.E. Hinton and R.R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” *Science* **313**, 504–507 (2006).
 - [10] H. W. Lin and M. Tegmark, “Why does deep and cheap learning work so well?” *ArXiv e-prints* **abs/1608.08225** (2016).
 - [11] G. Carleo and M. Troyer, “Solving the Quantum Many-Body Problem with Artificial Neural Networks,” *Science* **355**, 602–606 (2017).
 - [12] Kenneth G. Wilson, “The renormalization group: Critical phenomena and the kondo problem,” *Rev. Mod. Phys.* **47**, 773–840 (1975).
 - [13] H. David Politzer, “Reliable perturbative results for strong interactions?” *Phys. Rev. Lett.* **30**, 1346–1349 (1973).
 - [14] V. L. Berezinskii, “Destruction of Long-range Order in One-dimensional and Two-dimensional Systems having a Continuous Symmetry Group I. Classical Systems,” *Soviet Journal of Experimental and Theoretical Physics* **32**, 493 (1971).
 - [15] J.M. Kosterlitz and D. Thouless, “Ordering, metastability and phase transitions in two-dimensional systems,” *Journal of Physics C: Solid State Physics* **6**, 1181 (1973).
 - [16] L. P. Kadanoff, “Scaling laws for Ising models near $T(c)$,” *Physics* **2**, 263–272 (1966).
 - [17] Christof Wetterich, “Exact evolution equation for the effective potential,” *Physics Letters B* **301**, 90 – 94 (1993).
 - [18] Steven R. White, “Density matrix formulation for quantum renormalization groups,” *Phys. Rev. Lett.* **69**, 2863–2866 (1992).
 - [19] Shang-keng Ma, Chandan Dasgupta, and Chin-kun Hu, “Random antiferromagnetic chain,” *Phys. Rev. Lett.* **43**, 1434–1437 (1979).
 - [20] Philippe Corboz and Frederic Mila, “Tensor network study of the shastry-sutherland model in zero magnetic field,” *Phys. Rev. B* **87**, 115144 (2013).
 - [21] Sylvain Capponi, V. Ravi Chandra, Assa Auerbach, and Marvin Weinstein, “ p_6 chiral resonating valence bonds in the kagome antiferromagnet,” *Phys. Rev. B* **87**, 161118 (2013).
 - [22] A. Auerbach, *Interacting electrons and quantum magnetism* (Springer, 1994).
 - [23] Jose Gaite and Denjoe O’Connor, “Field theory entropy, the h theorem, and the renormalization group,” *Phys. Rev. D* **54**, 5163–5173 (1996).
 - [24] J. Preskill, “Quantum information and physics: some future directions,” *J. Mod. Opt.* **47**, 127–137 (2000).
 - [25] S.M. Apenko, “Information theory and renormalization group flows,” *Physica A* **391**, 62–77 (2012).
 - [26] B.B. Machta, R. Chachra, M.K. Transtrum, and J.P. Sethna, “Parameter space compression undelies emergent theories and predictive models,” *Science* **342**, 604–607 (2013).
 - [27] Cedric Beny and Tobias J Osborne, “The renormalization group via statistical inference,” *New Journal of Physics* **17** (2015).
 - [28] Jean-Marie Stephan, Stephen Inglis, Paul Fendley, and Roger G. Melko, “Geometric mutual information at classical critical points,” *Phys. Rev. Lett.* **112**, 127204 (2014).
 - [29] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” *ArXiv Physics e-prints* **physics/0004057** (2000).
 - [30] Hinton G.E., “Training Products of Experts by Minimizing Contrastive Divergence,” *Neural Computation* **14**, 1771–1800 (2002).
 - [31] E. Fradkin, *Field theories of Condensed Matter Physics* (Cambridge University Press, 2013).
 - [32] Michael E. Fisher and John Stephenson, “Statistical mechanics of dimers on a plane lattice. ii. dimer correlations and monomers,” *Phys. Rev.* **132**, 1411–1431 (1963).
 - [33] S.S. Schoenholz, E.D. Cubuk, D.M. Sussman, E. Kaxiras, and A.J. Liu, “A structural approach to relaxation in glassy liquids,” *Nature Physics* **12**, 469–471 (2016).
 - [34] M.I. Jordan and T.M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science* **349**, 255–260 (2015).
 - [35] Noam Slonim and Naftali Tishby, “Document clustering using word clusters via the information bottleneck method,” in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’00 (ACM, 2000) pp. 208–215.
 - [36] S. Haykin, *Neural Networks and Learning Machines* (Pearson, 2009).
 - [37] Theano Development Team, “Theano: A Python framework for fast computation of mathematical expressions,” *arXiv e-prints* **abs/1605.02688** (2016).
 - [38] P. Mehta and D. J. Schwab, “An exact mapping between the Variational Renormalization Group and Deep Learning,” *ArXiv e-prints* **abs/1410.3831** (2014).

Supplemental materials: Methods

ESTIMATING MUTUAL INFORMATION

Here we gather the notations and define formally the quantities appearing in the text. We derive in detail the expression for the approximate mutual information measure, which is evaluated numerically by the RSMI algorithm. This measure is given in terms of a number of probability distributions, accessible via Monte Carlo samples and approximated by contrastive divergence (CD) trained RBMs, or directly defined by (different) RBMs [36].

We consider a statistical system $\mathcal{X} = \{x_1, \dots, x_N\} \equiv \{x_i\}_{i=1}^N$ of classical Ising variables $x_i \in \{0, 1\}$, which can equally well describe the presence or absence of a dimer. The system \mathcal{X} is divided into a small “visible” area \mathcal{V} , an “environment” \mathcal{E} , and a “buffer” \mathcal{B} separating the degrees of freedom in \mathcal{V} and \mathcal{E} spatially (for clarity of exposition we assumed $\mathcal{B} = \emptyset$ in the main text). We additionally consider a set of “hidden” Ising variables $\mathcal{H} = \{h_1, \dots, h_{N_h}\}$. For the main RSMI network described in the text, \mathcal{H} has the interpretation of coarse-grained degrees of freedom extracted from \mathcal{V} .

We assume the data distribution $P(\mathcal{X}) = P(\mathcal{V}, \mathcal{B}, \mathcal{E})$ – formally a Boltzmann equilibrium distribution defined by a Hamiltonian $H(\{x_i\})$ – is *given* to us indirectly via random (Monte Carlo) samples of $(\mathcal{V}, \mathcal{B}, \mathcal{E})_i$. The distributions $P(\mathcal{V}, \mathcal{E})$ and $P(\mathcal{V})$ are defined as marginalizations of $P(\mathcal{X})$. Performing the marginalizations explicitly is computationally costly and therefore it is much more efficient to approximate $P(\mathcal{V}, \mathcal{E})$ and $P(\mathcal{V})$ using two RBMs of the type defined in and above Eq. (2), trained using the CD-algorithm [30] on the restrictions of $(\mathcal{V}, \mathcal{B}, \mathcal{E})_i$ samples. The trained networks, with parameters $\Theta^{(\mathcal{V}, \mathcal{E})}$ and $\Theta^{(\mathcal{V})}$ (which we refer to as Θ -RBMs) define probability distributions $P_\Theta(\mathcal{V}, \mathcal{E})$ and $P_\Theta(\mathcal{V})$, respectively [see also Fig. 1(B)]. From mathematical standpoint, contrastive divergence is based on minimizing a proxy to the Kullback-Leibler divergence between $P_\Theta(\mathcal{V}, \mathcal{E})$ and $P_\Theta(\mathcal{V})$ and the data probability distributions $P(\mathcal{V}, \mathcal{E})$ and $P(\mathcal{V})$, respectively, i.e. the training produces RBMs which model the data well [30].

The conditional probability distribution $P_\Lambda(\mathcal{H}|\mathcal{V})$ is *defined* by another RBM, denoted henceforth by Λ -RBM, with tunable parameters $\Lambda = (a_i, b_j, \lambda_i^j)$:

$$P_\Lambda(\mathcal{H}|\mathcal{V}) = \frac{e^{-E_\Lambda(\mathcal{V}, \mathcal{H})}}{\sum_{\mathcal{H}} e^{-E_\Lambda(\mathcal{V}, \mathcal{H})}}, \quad (5)$$

$$E_\Lambda(\mathcal{V}, \mathcal{H}) = \sum_{ij} -v_i \lambda_i^j h_j - \sum_i a_i v_i - \sum_j b_j h_j$$

In contrast to Θ -RBMs it will *not* be trained using CD-algorithm, since its objective *is not* to approximate the data probability distribution. Instead, the parameters Λ will be chosen so as to maximize a measure of mutual information between \mathcal{V} and \mathcal{E} . The reason for exclusion of a buffer \mathcal{B} , generally of linear extent comparable to \mathcal{V} , is that otherwise MI would take into account correlations of \mathcal{V} with its immediate vicinity, which are equivalent with short-ranged correlations within \mathcal{V} itself. We now derive the MI expression explicitly.

Using $P(\mathcal{V}, \mathcal{E})$ and $P_\Lambda(\mathcal{H}|\mathcal{V})$ we can define the joint probability distribution $P_\Lambda(\mathcal{V}, \mathcal{E}, \mathcal{H}) = P(\mathcal{V}, \mathcal{E})P_\Lambda(\mathcal{H}|\mathcal{V})$ and marginalize over \mathcal{V} to obtain $P_\Lambda(\mathcal{E}, \mathcal{H})$. We can then define the mutual information (MI) between \mathcal{E} and \mathcal{H} in the standard fashion:

$$I_\Lambda(\mathcal{H} : \mathcal{E}) = \sum_{\mathcal{H}, \mathcal{E}} P_\Lambda(\mathcal{E}, \mathcal{H}) \log \left(\frac{P_\Lambda(\mathcal{E}, \mathcal{H})}{P_\Lambda(\mathcal{H})P(\mathcal{E})} \right) \quad (6)$$

The main task is to find the set of parameters Λ which maximizes $I_\Lambda(\mathcal{H} : \mathcal{E})$ given the samples $(\mathcal{V}, \mathcal{E})_i$. Since $P(\mathcal{E})$ is not a function of Λ one can optimize a simpler quantity:

$$A_\Lambda(\mathcal{H} : \mathcal{E}) = \sum_{\mathcal{H}, \mathcal{E}} P_\Lambda(\mathcal{E}, \mathcal{H}) \log \left(\frac{P_\Lambda(\mathcal{E}, \mathcal{H})}{P_\Lambda(\mathcal{H})} \right) \quad (7)$$

Using the Θ -RBM approximations of the data probability distributions as well as the definition of the $P_\Lambda(\mathcal{H}, \mathcal{E})$ one can further rewrite this as:

$$A_\Lambda(\mathcal{H} : \mathcal{E}) = \sum_{\mathcal{H}, \mathcal{E}} P_\Lambda(\mathcal{E}, \mathcal{H}) \log \left(\frac{\sum_{\mathcal{V}} P_\Lambda(\mathcal{V}, \mathcal{H}) P_\Theta(\mathcal{V}, \mathcal{E}) / P_\Lambda(\mathcal{V})}{\sum_{\mathcal{V}'} P_\Lambda(\mathcal{V}', \mathcal{H}) P_\Theta(\mathcal{V}') / P_\Lambda(\mathcal{V}')} \right) \quad (8)$$

The daunting looking argument of the logarithm can in fact be cast in a simple form, using the fact that all the probability distributions involved either are of Boltzmann form, or marginalization thereof over the hidden variables,

which can be performed explicitly:

$$A_\Lambda(\mathcal{H} : \mathcal{E}) \equiv \sum_{\mathcal{H}, \mathcal{E}} P_\Lambda(\mathcal{E}, \mathcal{H}) \log \left(\frac{\sum_{\mathcal{V}} e^{-E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{E}, \mathcal{H})}}{\sum_{\mathcal{V}'} e^{-E_{\Lambda, \Theta}(\mathcal{V}', \mathcal{H})}} \right), \quad (9)$$

where

$$\begin{aligned} E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{E}, \mathcal{H}) &= E_\Lambda(\mathcal{V}, \mathcal{H}) + E_\Theta(\mathcal{V}, \mathcal{E}) + \sum_j \log[1 + \exp(\sum_i v_j \lambda_i^j + b_j)] \\ E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{H}) &= E_\Lambda(\mathcal{V}, \mathcal{H}) + E_\Theta(\mathcal{V}) + \sum_j \log[1 + \exp(\sum_i v_j \lambda_i^j + b_j)], \end{aligned} \quad (10)$$

and where $E_\Theta(\mathcal{V}, \mathcal{E})$ and $E_\Theta(\mathcal{V})$ are defined by the parameter sets $\Theta^{(\mathcal{V}, \mathcal{E})}$ and $\Theta^{(\mathcal{V})}$ of the trained Θ -RBMs:

$$\begin{aligned} P_\Theta(\mathcal{V}) &\propto e^{-E_\Theta(\mathcal{V})} \\ P_\Theta(\mathcal{V}, \mathcal{E}) &\propto e^{-E_\Theta(\mathcal{V}, \mathcal{E})} \end{aligned} \quad (11)$$

Note, that since in $P_\Lambda(\mathcal{H}|\mathcal{V})$ the a_i parameter dependence cancels out [and consequently also in $P_\Lambda(\mathcal{E}, \mathcal{H})$], the quantity A_Λ does not depend on a_i . Hence, without loss of generality, we put $a_i \equiv 0$ in our numerical simulations, i.e. the Λ -RBM is specified by the set of parameters $\Lambda = (b_j, \lambda_i^j)$ only.

A_Λ is an average over the distribution $P_\Lambda(\mathcal{E}, \mathcal{H})$ of a logarithmic expression [see Eq. (5)], which itself can be further rewritten as a statistical expectation value for a system with energy $E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{H})$, with variables \mathcal{H} held fixed:

$$\log \left(\frac{\sum_{\mathcal{V}} e^{-E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{E}, \mathcal{H})}}{\sum_{\mathcal{V}'} e^{-E_{\Lambda, \Theta}(\mathcal{V}', \mathcal{H})}} \right) = \log \left(\frac{\sum_{\mathcal{V}} e^{-E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{H}) - \Delta E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{E}, \mathcal{H})}}{\sum_{\mathcal{V}'} e^{-E_{\Lambda, \Theta}(\mathcal{V}', \mathcal{H})}} \right) \quad (12)$$

$$\equiv \log \left(\left\langle e^{-\Delta E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{E}, \mathcal{H})} \right\rangle_{\mathcal{H}} \right) \approx \langle -\Delta E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{E}, \mathcal{H}) \rangle_{\mathcal{H}} \quad (13)$$

with $\Delta E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{E}, \mathcal{H}) = E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{E}, \mathcal{H}) - E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{H})$. Thus finally, we arrive at a simple expression for A_Λ :

$$A_\Lambda(\mathcal{H} : \mathcal{E}) \approx \sum_{\mathcal{H}, \mathcal{E}} P_\Lambda(\mathcal{E}, \mathcal{H}) \langle -\Delta E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{E}, \mathcal{H}) \rangle_{\mathcal{H}}. \quad (14)$$

This expression can be numerically evaluated: using the fact that $P_\Lambda(\mathcal{E}, \mathcal{H}) = \sum_{\mathcal{V}'} P_\Lambda(\mathcal{H}|\mathcal{V}') P(\mathcal{V}', \mathcal{E})$ we replace the sums over \mathcal{V}' and \mathcal{E} with a Monte Carlo (MC) average over $N_{(\mathcal{V}, \mathcal{E})}$ samples $(\mathcal{V}', \mathcal{E})_i$. Furthermore, given a Λ -RBM (at current stage of training) and a sample of $(\mathcal{V})_i$, one can easily draw a sample $(\mathcal{H})_i \equiv (\mathcal{H}(\mathcal{V}))_i$ according to probability distribution $P_\Lambda(\mathcal{H}|\mathcal{V})$. Hence we have a MC estimate:

$$A_\Lambda(\mathcal{H} : \mathcal{E}) \approx \frac{1}{N_{(\mathcal{V}, \mathcal{E})}} \sum_{(\mathcal{V}', \mathcal{E}, \mathcal{H}(\mathcal{V}'))_i} \langle -\Delta E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{E}, \mathcal{H}) \rangle_{\mathcal{H}}. \quad (15)$$

The expectation value in the summand is itself also evaluated by MC averaging, this time with respect to Boltzmann probability distribution with energy $E_{\Lambda, \Theta}(\mathcal{V}, \mathcal{H})$.

Maximizing A_Λ with respect to parameters $\Lambda = (a_i, b_j, \lambda_i^j)$ is most easily achieved using conventional stochastic gradient descent procedure [36]. To this end we estimate the derivative $\partial_{\lambda_{ij}} A_\Lambda$ over samples $(\mathcal{V}, \mathcal{E}, \mathcal{H}(\mathcal{V}))_i$. More accurately, we divide the samples into mini-batches, obtain an average assessment of $\partial_{\lambda_{ij}} A_\Lambda$, and use it to update the parameters of the Λ -RBM: $\lambda_{ij}^j \rightarrow \lambda_{ij}^j - \eta \cdot \partial_{\lambda_{ij}} A_\Lambda$ (and similarly for b_j) with a learning rate η . This is then repeated for next mini-batch. A run through all mini-batches constitutes one epoch of training. In Fig. 1 of the Supplementary Materials we show the convergence of the A_Λ estimation and the development of the weight matrices for the case of the Ising system.

A practical remark is that the gradients should best be computed explicitly (a simple, if tedious, computation) prior to numerical evaluation, and one should not use the automated gradient computation capability provided by *e.g.* Theano package [37]. The reason is that some of the dependence on parameters Λ is stochastic (i.e. they define the probability distribution which the MC-averaging approximates), and this dependence may possibly not be captured correctly by automated gradient computing procedures.

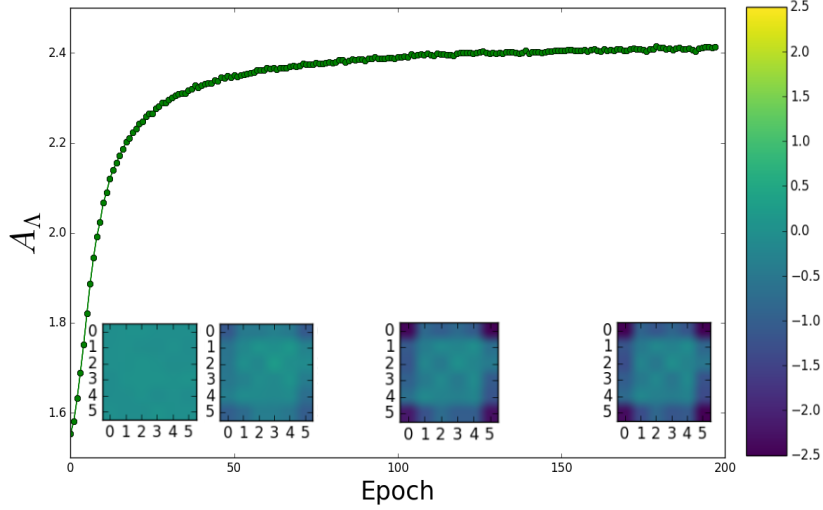


FIG. 5. The proxy A_Λ of mutual information, as a function of training epochs of the Λ -RBM for the Ising model data. The saturation behaviour of A_Λ demonstrates the convergence of the algorithm. The weight matrices shown are for a single hidden neuron, with visible area \mathcal{V} of size 6×6 , after 0, 10, 100 and 190 training epochs respectively. The development of the boundary coupling behaviour for Ising system, discussed in the main text and in the supplemental materials, can be seen.

RELATION OF MUTUAL INFORMATION RG PROCEDURE TO CONVENTIONAL RG SCHEMES

Here we provide an intuitive theoretical argument elucidating the connection of our mutual information based approach to more standard treatments of real-space RG. Information-theoretic approaches were also advocated or investigated in Refs. [23–27]. Before defining our explicit criteria for identifying relevant degrees of freedom, let us first briefly rephrase the conventional RG procedure in probabilistic terms.

Consider then a physical system, represented by a set of local degrees of freedom (or random variables) $\mathcal{X} = \{x_i\}$ and governed by a Hamiltonian energy function $H(\{x_i\})$. The equilibrium probability distribution is of Boltzmann form: $P(\{x_i\}) \propto e^{-\beta H(\{x_i\})}$, with β the inverse temperature. Next we consider a new and smaller set of degrees of freedom $\mathcal{H} = \{h_i\}$, i.e. the coarse-grained variables, whose dependence on \mathcal{X} is given by a conditional probability $\Pi_i P_\Lambda(h_i|\mathcal{X})$, where Λ are variational internal parameters to be specified and each h_i depends on some localized set of $\{x_i\}$. The RG transformation in this language consists of finding the effective Hamiltonian of the coarse-grained degrees of freedom $\tilde{H}(\mathcal{H})$ by marginalizing over (or integrating-out, in physical terms) degrees of freedom \mathcal{X} in the joint probability distribution of \mathcal{X} and \mathcal{H} :

$$\begin{aligned} \tilde{H} &= -\log(Z[\mathcal{H}]), \\ Z[\mathcal{H}] &= \sum_{\mathcal{X}} \Pi_i P_\Lambda(h_i|\mathcal{X}) e^{-\beta H}. \end{aligned} \quad (16)$$

The new effective energy function contains all the information required to evaluate expectation values of the h_i variables in an exact fashion.

The usefulness (and practicality) of the RG procedure depends on choosing $P_\Lambda(h_i|\mathcal{X})$ (or equivalently the relevant degrees of freedom) such that effective Hamiltonian \tilde{H} remains as short range as possible and (if \mathcal{H} is continuous) the fluctuations of \mathcal{H} are as small as possible, so that high powers of h_i are not needed. More formally we demand that the Taylor expansion of \tilde{H} in h_i :

$$\tilde{H} = \sum_i f_i h_i + \sum_{\langle ij \rangle} f_{ij} h_i h_j + \sum_{\langle\langle ij \rangle\rangle} f_{ij} h_i h_j + \sum_{\langle ijk \rangle} f_{ijk} h_i h_j h_k + \sum_{\langle\langle ijk \rangle\rangle} \dots \quad (17)$$

contains only short-ranged and few-body terms, i.e. the coefficients f decay exponentially with distance. Since the f coefficients can be computed as the cumulants, or connected n -point correlation functions, of the \mathcal{X} degrees of freedom in this theory, one can also rephrase the requirements: integrating out some subset of h_j must, via the coupling induced by parameters Λ , generate strong mass terms for the x_i which gap out all their long-range fluctuations. The correlation functions of x_i decay then exponentially and thus \tilde{H} is short-ranged. If the above requirements are satisfied, all the

terms in \tilde{H} beyond some finite distance can be removed while making only minor changes to statistical properties of system \mathcal{H} . The procedure can then be repeated recursively granting access to increasingly long-ranged features without keeping track of all degrees of freedom.

What constitutes a good RG scheme in the above language is intuitively clear, but hard to formalise, especially with an algorithmic goal in mind. The mutual information prescription, on the other hand, has a very precise formulation, and lends itself naturally to computational implementation. How are they related? For simplicity consider a one dimensional system \mathcal{X} , divided into three parts \mathcal{V}_a , \mathcal{V}_b , and \mathcal{V}_c . Denote the hiddens in these region by h_i^a , h_j^b , and h_k^c , respectively. Assume first that we have enough hiddens in these regions to maximize the mutual information with their complements, i.e. that adding any additional hiddens results in no further gains to MI. A term of the type $f_{ijk} h_i^a h_j^b h_k^c$ in the effective Hamiltonian implies that the probability distribution of h^a depends directly on h^b and h^c and not just on h^b . Since the underlying $P(\mathcal{X})$ distribution is local, such a dependence can only be mediated by degrees of freedom in \mathcal{V}_b which are correlated with degrees of freedom in \mathcal{V}_c . However, by our assumption of maximizing the mutual information, h_j^b already couple to all such degrees of freedom in \mathcal{V}_c . Consequently, such a term cannot exist when the mutual information is maximal. Thus, as MI grows, the coefficients of longer-ranged terms in \tilde{H} should decay.

The argument above justifies why the MI-based RG procedure should give results similar to more conventional RG schemes. This is explicitly confirmed by our numerical results for the dimer and Ising models. A formal, analytical investigation of this relation is nevertheless desirable, and will be pursued in a separate work.

UNIFORM VS. BOUNDARY CENTERED FILTERS FOR ISING MODEL

We comment briefly on the numerical results obtained for the Ising model, namely the boundary coupling behaviour observed in the weights λ_i^j for increasing sizes of the visible area \mathcal{V} [see Fig. 2(E) in the main text]. As mentioned in the main text, this behaviour, generalizing the simple Kadanoff blocking seen for 2×2 areas \mathcal{V} , can in fact be justified on the grounds of conventional (in the sense defined in the previous section) RG scheme behaviour, further validating the correctness of our numerical procedure.

Let us then state a concrete question: should the parameters λ_i^j defining slow degrees of freedom in a real-space RG scheme [as described in the previous section of Supplemental Materials around Eqs. (12,13)], in the limit of a large area \mathcal{V} , have a uniform modulation (pattern) over the area \mathcal{V} or should they more strongly couple to the boundary of \mathcal{V} ? We argue that for the Ising case it is the latter, therefore the results obtained by our MI-based procedure are consistent with behaviour expected of an RG scheme. For the sake of argument we assume the simplest scenario, when the system has only short range interactions and the number of “hiddens” coupling to \mathcal{V} equal the number of degrees of freedom on its boundary $\partial\mathcal{V}$ (defined as the set of degrees of freedom in \mathcal{V} with neighbours outside \mathcal{V}). It is easy to see that in this case the optimal solution is to couple each hidden unit h_i , with infinite strength, to one degree of freedom v_i in $\partial\mathcal{V}$, effectively enforcing $h_i \equiv v_i$ there. In the spirit of the discussion in the previous section of Supplemental Materials, integrating out the h_i would generate large mass terms for the v_i on the boundary, removing all correlations between the inside and outside of \mathcal{V} and consequently the effective Hamiltonian for the h_i would be short-ranged. For number of hidden units smaller than $\partial\mathcal{V}$, the filters would keep favoring the boundary over the bulk of the cell to keep track of the degrees of freedom most directly coupled to the environment, an intuition which is qualitatively confirmed by our numerical results for the Ising model.

For the case of the dimer system the weight matrices appear uniformly textured even for a large area \mathcal{V} [see Fig. 4(A) in the main text], which may seem puzzling in light of the above argument. This is the result of dimers being constrained degrees of freedom (or, in the continuum formulation, the height field obeying a conservation equation) [31], as opposed to Ising spins. The height field h is not a microscopically accessible quantity in dimer model, only its gradient ∇h is. The weights λ_i^j we find for the dimer model enforce a uniform coupling pattern of the hidden degrees of freedom to the gradient of height field over many unit cells contained in the visible area \mathcal{V} , i.e. the hiddens couple to an average of the electric field in \mathcal{V} . This integral of ∇h over \mathcal{V} is, however, equivalent to a difference of boundary terms for the height field h itself! Thus there is no contradiction: the direct coupling of hiddens to the height field on the boundary, which is physically not possible for the dimer model, is achieved by a uniform coupling to its gradient.

In principle, the type of scaling behaviour under the size of \mathcal{V} may allow to draw conclusions as to whether the degrees of freedom are constrained or not.

COMPARISON WITH CONTRASTIVE DIVERGENCE TRAINED RBMS

The relation between unsupervised machine learning (in particular in deep NNs), and RG, has been a subject of some controversy recently. Ref.[38] claims an “exact” mapping between the two, while Tegmark and Lin assert, to the

contrary, that RG is entirely unrelated to unsupervised learning [10]. Our theoretical arguments, as well as explicit numerical results on dimer model disprove the very general claims of Ref. [38], however, we cannot entirely agree with Tegmark and Lin either.

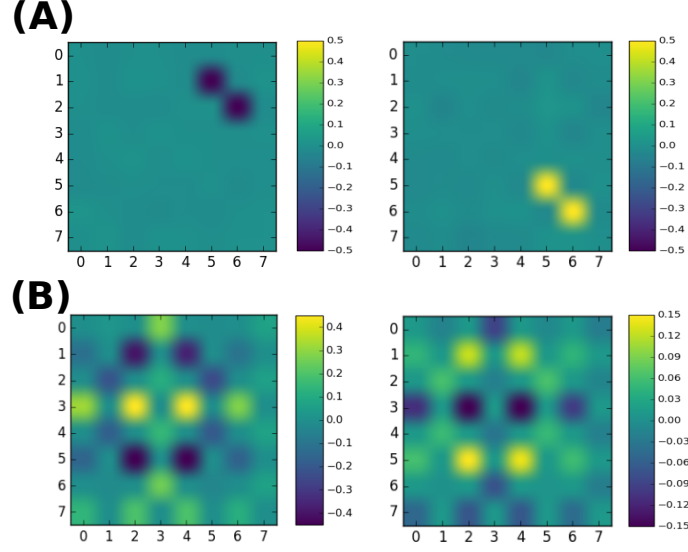


FIG. 6. The weights of a standard CD-trained RBM on dimer model with additional noise: (A) For a small number of hidden neurons the network learns to recognize the spin-pair pattern (B) Given enough filters to capture all the spin pairs, the successive filters capture dimer configurations more similar to the “columnar” state in right panel of Fig. 3(B), instead of the staggered configurations the RSMI finds.

The crucial point is that a neural network is not specified without the cost function. Both Ref. [38] (explicitly) and Ref. [10] (implicitly) assume that the networks are trained using Kullback-Leibler divergence, or some related measure, resulting in the network which approximates well the input data probability distribution. Ref. [10] is correct in pointing out that this will not generally result in RG transformation. We are not limited, however, to such cost functions; in point of fact, we are free to choose a cost function with an entirely different objective. As we have shown, choosing to maximize the mutual information with the system at large results in a network identifying the physically relevant degrees of freedom.

To disprove explicitly that a generic NN trained to recognize patterns performs a physically meaningful coarse-graining (i.e., potentially, RG), we examine the weight matrices of CD-trained network (Θ -RBM) on the dimer model with additional spin “noise”, the same we considered in the main text. In Fig. 6(A) we show the examples of weights obtained for a small number of hidden units: the network strongly couples to individual pairs of spins fluctuating in sync, even though they are irrelevant from the point of view of physics. This is correct behaviour, when patterns are to be discerned (since there are many entirely different dimer textures, but the same fluctuating pattern of spin pairs for each configuration), but does not make any sense physically. If given a few hidden units we were to extract new degrees of freedom to further continue the RG procedure on, we would discard *all of* the dimer model in the first step, ending up with a trivial system of decoupled spin pairs. Only given enough hidden neurons to capture all the spin pairs does the CD-trained network learn to discern extensive dimer textures, and even then they are not the staggered configurations giving electric fields important from the point of RG, but rather more similar to the high-entropy columnar configurations (which map to vanishing electric fields).